Complex Selection on Human Polyadenylation Signals Revealed by Polymorphism and Divergence Data

Yaroslav A. Kainov[1,2,*,†], Vasily N. Aushev[2,3,†], Sergey A. Naumenko[4,5], Elena M. Tchevkina[2], and Georgii A. Bazykin[4,6,7,8,9,*]

[1]Centre for Developmental Neurobiology, King's College London, London, United Kingdom

[2]Oncogenes Regulation Department, N.N. Blokhin Russian Cancer Research Center, Institute of Carcinogenesis, Moscow, Russia

[3]Department of Preventive Medicine, Icahn School of Medicine at Mount Sinai, New York

[4]Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences, Moscow, Russia

[5]Genetics and Genome Biology Program, The Hospital for Sick Children, Toronto, Canada

[6]Skolkovo Institute of Science and Technology, Skolkovo, Russia

[7]Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Russia

[8]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Russia

[9]Pirogov Russian National Research Medical University, Moscow, Russia

*Corresponding author: E-mail: yaroslav.kainov@kcl.ac.uk; gbazykin@iitp.ru.

[†]These authors contributed equally to this work.

Accepted: June 5, 2016

Abstract

Polyadenylation is a step of mRNA processing which is crucial for its expression and stability. The major polyadenylation signal (PAS) represents a nucleotide hexamer that adheres to the AATAAA consensus sequence. Over a half of human genes have multiple cleavage and polyadenylation sites, resulting in a great diversity of transcripts differing in function, stability, and translational activity. Here, we use available whole-genome human polymorphism data together with data on interspecies divergence to study the patterns of selection acting on PAS hexamers. Common variants of PAS hexamers are depleted of single nucleotide polymorphisms (SNPs), and SNPs within PAS hexamers have a reduced derived allele frequency (DAF) and increased conservation, indicating prevalent negative selection; at the same time, the SNPs that "improve" the PAS (i.e., those leading to higher cleavage efficiency) have increased DAF, compared to those that "impair" it. SNPs are rarer at PAS of "unique" polyadenylation sites (one site per gene); among alternative polyadenylation sites, at the distal PAS and at exonic PAS. Similar trends were observed in DAFs and divergence between species of placental mammals. Thus, selection permits PAS mutations mainly at redundant and/or weakly functional PAS. Nevertheless, a fraction of the SNPs at PAS hexamers likely affect gene functions; in particular, some of the observed SNPs are associated with disease.

Key words: polyadenylation, AATAAA, 1000 genomes, SNP, mRNA processing.

Introduction

Polyadenylation is an essential step of mRNA processing in eukaryotes. It affects many aspects of mRNA physiology and plays an important role in its dynamics. Over 50% of human genes contain more than one potential site of cleavage and polyadenylation (Tian et al. 2005; Shepard et al. 2011). A process called alternative polyadenylation (APA) leads to generation of mRNA isoforms with different lengths of 3′-untranslated region or even truncated protein-coding regions (di Giammartino et al. 2011). The pattern of mRNA polyadenylation undergoes dramatic changes during cell differentiation, proliferation, and malignant transformation (Sandberg et al. 2008; Ji et al. 2009; Singh et al. 2009).

Polyadenylation includes two major steps: recognition of polyadenylation signals (PAS) leading to mRNA cleavage and nonmatrix addition of polyA tail (Colgan and Manley 1997).

Polyadenylation is a complex process regulated by a variety of trans-acting protein factors and cis-elements of mRNA. mRNA 3′-processing complex contains up to 85 proteins (Shi et al. 2009) including CPSF (cleavage and polyadenylation specificity factor), a multisubunit complex which plays a crucial role in mRNA cleavage and polyadenylation. CPSF binds a specific common PAS, an AATAAA hexamer (or its close variant) usually located within 50 nucleotides upstream of the cleavage site (Chan et al. 2014). PAS is present in almost 90% of mammalian mRNAs, and is the most common and best studied signal of polyadenylation (Proudfoot 1991; Beaudoing et al. 2000; Tian et al. 2005; Cheng et al. 2006).

Polyadenylation is tuned by natural selection. Cleavage sites and patterns of their usage are conserved across mammals (Ara et al. 2006; Lee et al. 2008). The regions of 3′-UTRs carrying PAS hexamers are depleted of single nucleotide polymorphisms (SNPs) (Castle 2011), and their deletion or mutation leads to a dramatic decrease in expression of target mRNAs due to changes in polyadenylation (Yang et al. 2009; Nunes et al. 2010) and/or transcription efficiency (Mapendano et al. 2010). Additionally, mutations disrupting a PAS located near an alternative cleavage site or affecting its strength influence the site usage and might be clinically relevant (Thomas and Saetrom 2012). Therefore, PAS hexamers are expected to be strongly selected. These patterns of selection are informative of the functional significance of mutations and may help to improve clinical predictions of mutation effects (Adzhubei et al. 2010; Stecher et al. 2014). However, they have never been studied systematically. Selection may be estimated from data on divergence with related species, or from within-species polymorphism. Divergence data provide more power, as SNP densities are still lower than densities of interspecies substitutions. On the other hand, polymorphism is immune to interspecies changes of fitness landscapes, for example, situations when a mutation deleterious in one species is harmless in another (Kondrashov, Sunyaev et al. 2002; Kern and Kondrashov 2004; Mustonen and Lassig 2009; Naumenko et al. 2012).

The current avalanche of data on human population-level polymorphism allows measuring patterns of selection with unprecedented resolution. From a single genome, selection favoring or disfavoring a signal may be inferred from its genomic over- or underrepresentation, respectively. From polymorphism data, selection may be inferred from densities of SNPs or allelic frequencies within those SNPs. The dependence of the overall level of polymorphism on selection may be complex even in the simplest single-locus case (McVean and Charlesworth 2000; Kondrashov et al. 2006; Schmidt et al. 2008), and is further complicated by linkage between sites (Genomes Project et al. 2010; Wilkening et al. 2013; You et al. 2015). However, the situation is simplified when alleles may be a priori subdivided into preferred and unpreferred. Negative selection may then be inferred from underrepresentation of SNPs, or from reduced allele frequencies of such SNPs, at sites occupied by the favored allele, compared to a neutral control. Conversely, positive selection is manifested as an excess of SNPs, and increased allele frequencies of such SNPs, at sites of the disfavored allele, compared to a neutral control (although it simultaneously purges variation at linked sites). Here, we use the data on divergence between species of placental mammals and human polymorphism data of the 1000 Genomes Project (Genomes Project et al. 2010) to comprehensively analyze the patterns of selection acting on PAS hexamers.

Materials and Methods

Source Data Sets

Lists of cleavage site positions (polyAsite.db2), positions of PAS hexamers (PAS.db2) and gene identifiers (gene.db2) were obtained from PolyA.db2 database (Lee et al. 2007) (http://polya.umdnj.edu/polya_db/v2/download/, last accessed June 14, 2016). Genomic coordinates were converted from hg17 to hg19 version using liftOver tool from UCSC (https://genome.ucsc.edu/cgi-bin/hgLiftOver, last accessed June 14, 2016).

Polymorphisms data, including allele frequencies, from Interim Phase 1 of 1000 Genomes project (Genomes Project et al. 2012) were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ (last accessed June 14, 2016), last accessed June 14, 2016. This data set comprises the genotypes of 1,094 individuals, and includes a total of 37,852,169 autosomal SNPs. Only true SNPs, that is, those where the reference and alternative alleles differed in a single-nucleotide mismatch, were included in the analysis. PhastCons score (Siepel et al. 2005) data set for placental mammals was downloaded from UCSC server on February 2, 2016. OMIM data (Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University [Baltimore, MD]) were downloaded on October 15, 2014 from the omim.org FTP server as a plain text file. SNP identifiers were extracted from the text and queried for the intersection with the polymorphisms we found in PAS hexamers. ClinVar data were downloaded from the ClinVar catalogue FTP server (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/, last accessed June 14, 2016) (Landrum et al. 2014) on May 26, 2014. GWAS data were downloaded from NHGRI GWAS site (www.genome.gov/gwastudies, last accessed June 14, 2016) (Welter et al. 2014) on June 18, 2014. We analyzed the intersections between dbRS ids from GWAS and ClinVar databases and polymorphisms observed in PAS hexamers.

Retrieving Sequences of PAS Hexamers and Control Hexamers

Sequences of PAS hexamers were retrieved from the reference human genome (assembly hg19, GRCh37) according to the positions indicated in the PAS.db2 database. A small fraction (<1%) of PAS hexamers that did not match the reference

human genome was discarded. Each nucleotide observed at a SNP position was categorized as ancestral or derived, using the ancestral human genome sequence retrieved from Ensembl FTP server, and derived allele frequency (DAF) was measured as the fraction of genotypes carrying the derived allele. As a control, we selected hexamers located in the same 3′-UTR regions but on the opposite (noncoding) DNA strand, and not observed as a PAS in the PAS.db2 database. SNP density was defined as the ratio of the number of SNPs within hexamers to the total length of hexamers. The mean phastCons score for each PAS was extracted from the phastCons data set for placental mammals.

Each SNP was characterized as "improvement" if the derived hexamer ranked higher than the ancestral hexamer in the list of 13 hexamers sorted by genomic frequency; as "impairment" if it ranked lower than the ancestral one; and as "disruption" if it did not belong to this list.

When a PAS hexamer could not be annotated unambiguously, it was excluded from the corresponding comparison. The final set of characterized PAS hexamers is presented as supplementary table S3, Supplementary Material online.

Statistical Analysis

Differences in SNP densities and DAFs were compared using the two-tailed Fisher's exact test and two-tailed Mann–Whitney *U*-test, respectively. In comparisons of functional groups, the considered group was compared to the remainder of the sample. Statistical significance was defined as $P < 0.05$. All statistical tests were performed in R. Plots were created with ggplot2 R package.

Results

The PAS hexamers typically have one of the 13 nucleotide sequences. The ranking of these PAS hexamers according to their frequencies in the genome is consistent with their ranking according to their efficiency in stimulation of cleavage and polyadenylation (supplementary table S1, Supplementary Material online). In particular, the first two of these sequences—AATAAA and ATTAAA—are by far the most frequent, together comprising 55.4% of all hexamers in the human genome; and their efficiency (Sheets et al. 1990) substantially exceeds that of all lower-ranking hexamers (supplementary table S1, Supplementary Material online).

To measure selection, we analyze two aspects of polymorphism: SNP densities and within-population frequencies of nonancestral nucleotides (DAFs), as well as interspecies conservation. These three measurements provide complementary estimates of selection. As multiallelic SNPs are rare in this data set (Genomes Project et al. 2010), DAFs are expected to be only dependent on the strength of selection at corresponding sites. However, SNP densities and rate of sequence divergence also depend on the mutation rate. To study selection, we therefore compare the properties of PAS hexamers to those of control hexamers. Control hexamers were chosen so that they have the same nucleotide sequence, and are located in similar regions of 3′-UTRs, but are positioned on the opposite strand, and therefore cannot be functional PAS hexamers (see Materials and Methods). This approach controls for most sources of local as well as global nonuniformity of the mutation rates.

Among the 55,856 investigated PAS hexamers (an average of 3.1 hexamers per gene), 2,066 (3.7%) were polymorphic, according to The 1000 Genomes Project database (supplementary table S2, Supplementary Material online). and only 47 (2.3%) of them carried more than one SNP. As a whole, PAS hexamers were not depleted of SNPs, compared to the control sample ($P = 0.77$, two-tailed Fisher's exact test, fig. 1*A*), although the density of SNPs was reduced in hexamers AATAAA and ATTAAA (fig. 3*A*), and also in those PAS hexamers that were more likely to be functional (see below). However, DAFs of the observed SNPs were reduced at PAS hexamers compared to the control (fig. 1*B*), while interspecies conservation was higher than in the control (fig. 1*C*), indicating negative selection against such mutations.

Knowledge of the relative strength of different PAS hexamers allowed us to predict the effect of mutations on their efficiency. We categorized SNPs at PAS hexamers as "disrupting" if the derived hexamer was not one of the 13 legitimate PAS hexamer sequences. The remaining SNPs were categorized as "impairing" if they reduced the rank of the hexamer, or "improving," if they increased it.

Overall, we did not observe a measurable enrichment or depletion for any of these three classes of SNPs, compared to the control (supplementary table S2, Supplementary Material online); however, these SNPs differed in their DAFs. The disrupting SNPs segregated at lower DAFs than the control (fig. 1*D*), indicating negative selection against them. For impairing or improving SNPs, the difference in DAFs from the control was not significant, although on average the impairing SNPs had somewhat lower DAFs, while improving SNPs had higher DAFs than expected. The impairing and disrupting SNPs had significantly lower DAFs than improving SNPs in the PAS hexamers, but not in the control hexamers, indicating negative selection against disrupting and impairing mutations, and/or positive selection in favor of improving mutations in PAS hexamers (fig. 1*D*). Frequency spectrums for the examined SNPs are represented in supplementary materials (supplementary figs. S1–S3, Supplementary Material online).

Next, we stratified the PAS hexamers according to several characteristics, and analyzed the differences in SNP densities, allele frequencies, and interspecies conservation between the categories (fig. 2). A few patterns emerged with regard to the densities of SNPs in different classes (fig. 3*A*). First, each gene can have either one polyadenylation-associated cleavage site or multiple alternative sites (fig. 2*A*). SNP density was the lowest at the PAS hexamers corresponding to the only cleavage site in a gene ("unique"), and was higher if multiple

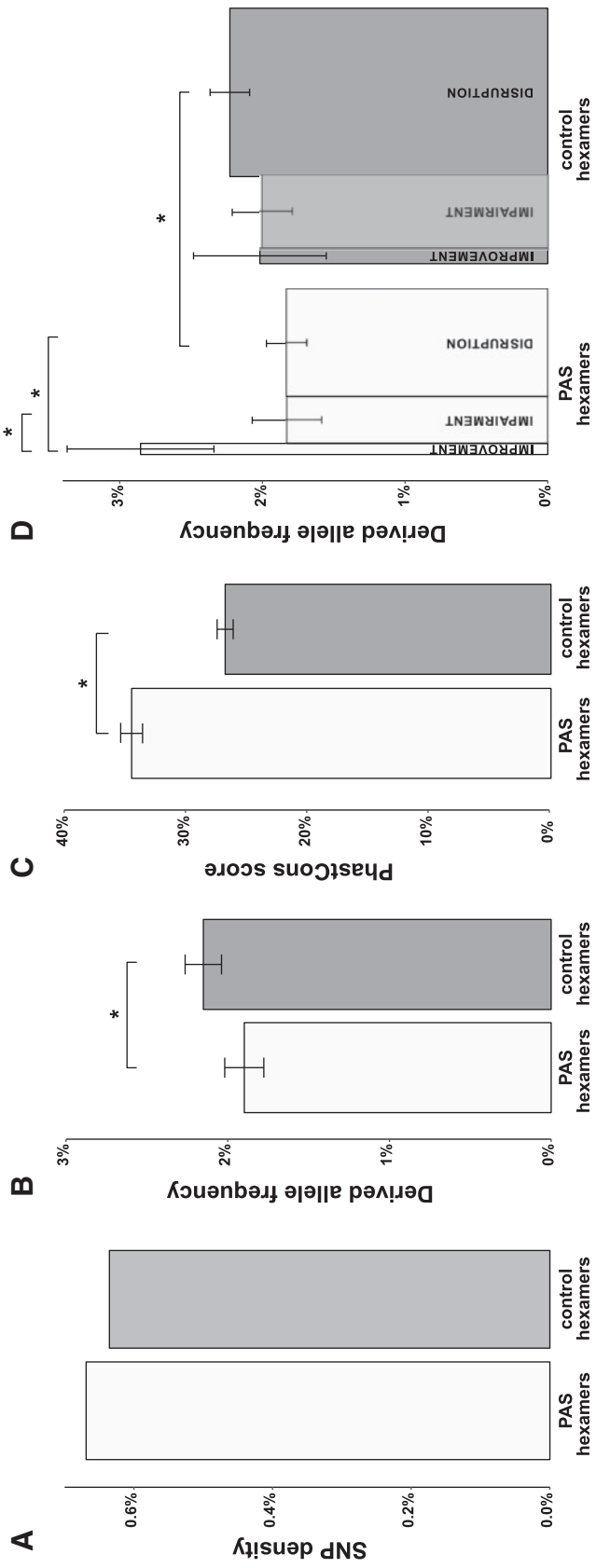


FIG. 1.—Patterns of selection in PAS hexamers. Whiskers represent standard errors of the mean. Asterisks correspond to $P < 0.05$ according to the Mann–Whitney's *U*-test. (*A*) Densities of SNPs in PAS hexamers and in the control sample. (*B*) Mean DAF of SNPs is reduced in PAS hexamers, compared to the control sample. (*C*) Mean phastCons score is increased in PAS hexamers, compared to the control sample. (*D*) DAF depends on the effect of SNPs on the functional activity of the PAS hexamer (box width represents the number of sites SNPs in the category).

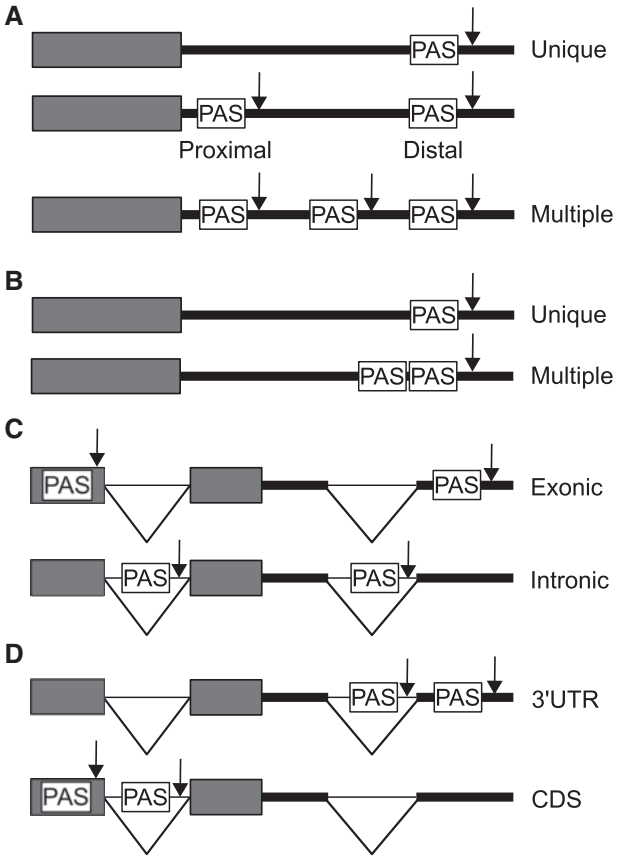


Fig. 2.—Schematic representation of functional classification of PAS hexamers. PAS hexamers are categorized according to the number and position of corresponding cleavage sites (*A*); number of PAS hexamers corresponding to a single cleavage site (*B*); localization within exon or intron (*C*); or localization within CDS or 3′-UTR (*D*). Gray boxes, coding exons; thick lines, 3′-UTR exons; angled lines, introns; arrows, cleavage sites.

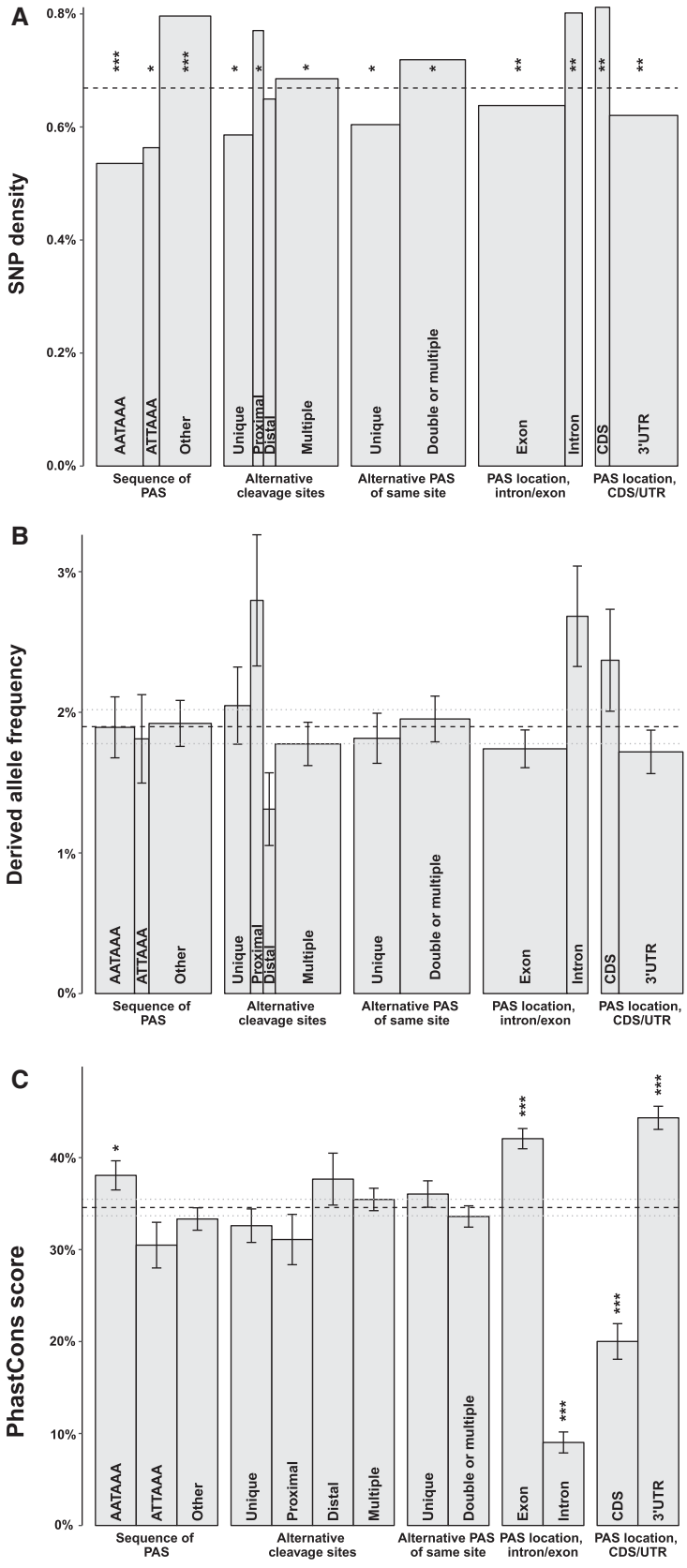
cleavage sites were present. When two cleavage sites were present, the PAS hexamer corresponding to the cleavage site distal from the promoter (i.e., located at the 3′-end of the longest mRNA isoform) had lower SNP densities. Second, several PAS hexamers may be associated with a single cleavage site (fig. 2*B*). At such sites, SNP densities were higher, compared to cleavage sites with unique PAS hexamers. Third, a fraction of PAS hexamers was located in an intron; such hexamers usually corresponded to alternative cleavage sites (fig. 2*C*). They had higher SNP densities than the exonic hexamers. Fourth, a fraction of PAS hexamers fell between the start and the stop codon; such hexamers are, of course, always alternative, and are usually intronic (fig. 2*D*). As expected, these hexamers were enriched in SNPs, compared with the PAS hexamers located within the 3′-UTRs. Interspecies conservation data demonstrated similar trends; specifically, the mean phastCons score was significantly higher for the "strongest" A ATAAA hexamer and for PAS-hexamers located in exons and 3′-UTRs (fig. 3*C*). While we saw no significant differences between investigated categories in DAFs (fig. 3*B*), the overall patterns were roughly coincident with those observed for SNPs densities and phastCons scores.

To elucidate the potential association of mutations at PAS hexamers with human diseases, we screened the OMIM (Online Mendelian Inheritance in Man), ClinVar (Clinical Variations), and NHGRI GWAS (The National Human Genome Research Institute Genome-Wide Association Studies) databases for PAS-affecting SNPs. We found five SNPs located within the PAS hexamers of five genes (table 1). Somewhat unexpectedly, all the observed mutations had rather high DAFs (> 1%) in the 1000 Genomes data set. Moreover, only two of the SNPs (rs78378222 and rs986475) affected PAS hexamers corresponding to unique polyadenylation sites (one per gene), whereas the other three SNPs affected the signals near alternative sites, two proximal and one distal. The observed SNPs also differed in their effect on polyadenylation site activity. While rs78378222 and rs986475 impair or disrupt the PAS, leading to a decrease of transcript polyadenylation and protein production (Delahaye et al. 2011; Stacey et al. 2011), rs10954213 improves the proximal PAS, resulting in an increased rate of formation of the shorter 3′-UTR isoform and higher mRNA stability and protein expression (Graham et al. 2007); and rs884205 represents an example of de novo formation of alternative PAS from the ancestral CTTA AA hexamer, which is not a PAS hexamer.

Discussion

In this work, we use the rich whole-genome data set on human polymorphism, The 1000 Genomes Data Set, and interspecies conservation to study the patterns of selection at PAS hexamers. We find that, overall, DAFs at PAS hexamers are lower, and conservation is higher, than in the control sequences, implying negative selection against mutations at PAS hexamers. We see no corresponding reduction in SNP densities, although SNP densities are reduced in the two strongest hexamers AATAAA and ATTAAA which together comprise over a half of the sample. This suggests that the typical selection at less strict PAS hexamers has a moderate strength, so that, although it is capable to reduce the frequency of the inferior allele and to prevent its fixation between species, it is seldom able to eliminate it completely.

However, the overall genome-wide patterns give only a crude understanding of selection. Categorization of alleles by their effect on the PAS hexamer efficiency reveals a more complicated picture. As expected, the DAFs of disrupting SNPs were substantially reduced, indicative of substantial negative selection against such mutations. The DAFs of the impairing SNPs were also reduced, compared with the control; in contrast, the DAFs of the improving SNPs were increased (fig. 1). Although the difference of the impairing and improving SNPs



from the control was statistically insignificant, they significantly differed from each other: the DAFs of impairing SNPs were significantly lower than those of improving SNPs. This implies that the impairing SNPs are negatively selected, that the improving SNPs are positively selected, or possibly both.

In the mutation-selection-drift balance, a continuous influx of deleterious mutations counteracted by selection against them leads to maintenance of an equilibrium concentration of suboptimal alleles in the genome. Under very weak selection, this may lead to alternating fixations of optimal and suboptimal alleles at a locus (Ohta's turnover) (Kimura and Ohta 1971; Ohta 1992; Charlesworth and Eyre-Walker 2007; Denisov et al. 2014). Here, we observe a different manifestation of the same phenomenon: a downward or upward bias in the mean frequency of the derived allele caused by negative and positive selection, respectively.

The strength and/or ubiquity of selection depend on the location of the PAS hexamer. It appears to be primarily determined by whether a mutation within a PAS hexamer may be circumvented by exploiting an alternative PAS hexamer. Mutations at unique PAS hexamers associated with unique cleavage sites are under the strongest selection, while the presence of an alternative PAS hexamer and/or cleavage site relaxes selection against the mutations. Specifically, the presence of another PAS hexamer reduced the action of selection both on the distal and the proximal ($P < 0.0001$, Fisher's test) PAS hexamer, compared to the unique ones. Overall, genomic redundancy tends to be associated with reduced selection against mutations within each functional element; for example, alternative splice sites and duplicate genes are under weaker selection than constitutive sites (Kurmangaliyev et al. 2013) and single-copy genes (Force et al. 1999; Kondrashov, Rogozin et al. 2002), respectively.

The distal PAS hexamers are under stronger selection than the proximal ones. The proximal hexamers tend to be further from the consensus sequence than the distal hexamers (Tian et al. 2005). This difference may facilitate proximal-to-distal cleavage site usage switching that occurs during a wide range of normal and pathological processes (Tian et al. 2005; di Giammartino et al. 2011). Thus, the less manifested consensus sequence of the proximal PAS hexamers could reduce the effect of the SNPs, compared with the distal PAS hexamers. Additionally, activity of proximal sites could be regulated by other polyadenylation factors (in particular, CSTF proteins)

Table 1
Characteristics of PAS Hexamers Carrying SNPs Associated with Human Pathologies

dbSNP ID	Gene Name	Normal PAS	Risk-Associated PAS	Frequency of Risk-Associated Allele	Type of PAS Cleavage Site	Effect of SNP on PAS	Phenotype	Database
rs78378222	*TP53*	AATAAA	AATACA	0.01	Unique	Impairment	Basal cell carcinoma	GWAS, ClinVar, OMIM
rs12721054	*APOC1*	AATAAA	AATGAA	0.03	Proximal	Impairment	High blood triglicerides	GWAS
rs10954213	*IRF5*	AATGAA	AATAAA	0.53	Proximal	Improvement	Systemic lupus erythematosus	ClinVar, OMIM
rs986475	*NCR3*	AATAAA	AACAAA	0.1	Unique	Disruption	Gastrointestinal stromal tumors	OMIM
rs884205	*TNFRSF11A*	CTTAAA	ATTAAA	0.19	Distal	De novo formation	Abnormal bone mineral density	GWAS

Fig. 3.—Polymorphism in different functional groups of PAS hexamers. (*A*) SNP densities; (*B*) DAFs; (*C*) PhastCons scores. In *A* and *B*, box width represents the number of SNPs in the group. Dashed lines represent the mean value in the entire sample, and dotted lines, its standard error. Whiskers represent standard error of the mean. Asterisks identify difference of the particular group from the remaining PAS hexamers, according to Fisher's exact test or Mann–Whitney *U*-test; $*P < 0.05$, $**P < 10^{-3}$, $***P < 10^{-10}$.

that interact with the downstream GU-rich sequence (Takagaki and Manley 1998; Nunes et al. 2010; Yao et al. 2012), making the nucleotide sequence of PAS hexamers less critical.

The exonic sites are under stronger selection than intronic sites. Also, the PAS hexamers located within the coding regions (which are typically also intronic) are under weaker selection than the 3′-UTR PAS hexamers (which tend to be exonic). This is intuitive, as the intronic sites are commonly alternative, whereas many of the exonic sites are constitutive. Additionally, intronic polyadenylation sites function in a tight cross-talk with splicing, and their usage could be regulated more or less independently of the binding of "classical" polyadenylation factors to the PAS hexamer. Thus, the expression level of splicing factors, which interact with specific signals independently of or even in competition with polyadenylation factors, and the strength of the 5′-splicing sites play an important role in regulation of the activity of intronic polyadenylation sites (Castelo-Branco et al. 2004; Kaida et al. 2010). Strikingly, almost all the observed differences between the functional groups of PAS hexamers coincide well with the trends of phylogenetic conservation of cleavage sites (Lee et al. 2008), supporting the key role of PAS hexamers in regulation of cleavage and polyadenylation.

Some of the SNPs that affect PAS hexamers might be associated with pathology. Interestingly, a fraction of these "pathological" SNPs affected alternative sites, suggesting that APA is important for physiological gene expression and function. Other germline and somatic mutations that affect alternative PAS hexamers have been previously described as implicated in pathogenesis of human type 1 diabetes, IPEX (immune dysfunction, polyendocrinopathy, enteropathy, X-linked), panic disorder (Bennett et al. 2001; Shin et al. 2007; Gyawali et al. 2010), and tumorigenesis (Wiestner et al. 2007).

In conclusion, the patterns of polymorphism within PAS hexamers reveal weak selection acting at these sites. This selection appears to be primarily determined by the direction and extent of the effect of the corresponding mutation on polyadenylation. While the destroying SNPs tend to be negatively selected, we find evidence of positive selection favoring the mutations that make the hexamers more similar to the consensus sequence, indicating that the nonconsensus sequences are mostly suboptimal. While SNPs are rare at those hexamers where they substantially disrupt the function, polymorphism within many of the rarely used hexamers appears to be nearly neutral. Pathogenic mutations may affect polyadenylation via a broad range of mechanisms, including disruption of existing, constitutive, or alternative, sites, improvement of an existing site, or even creation of a spurious site. Furthermore, the link between the changes in polyadenylation and the changes in expression is frequently nonlinear (Spies et al. 2013; Gupta et al. 2014). Annotation of possible effects of SNPs on polyadenylation should be included in any prediction of effects of both somatic and germline mutations; however, the complications listed above make such predictions inherently difficult.

Acknowledgments

The authors thank Eugene Makeyev and Alexej Abyzov for insightful comments. This work was supported by the Russian Foundation for Basic Research grant 15-34-21135 mol_a_ved and by the Molecular and Cellular Biology Program of the Russian Academy of Sciences.

Supplementary Material

Supplementary tables S1 and S2 and figures S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Literature Cited

Adzhubei IA, et al. 2010. A method and server for predicting damaging missense mutations. Nat Methods. 7:248–249.
Ara T, Lopez F, Ritchie W, Benech P, Gautheret D. 2006. Conservation of alternative polyadenylation patterns in mammalian genes. BMC Genomics 7:189.
Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. Genome Res. 10:1001–1010.
Bennett CL, et al. 2001. A rare polyadenylation signal mutation of the *FOXP3* gene (AAUAAA–>AAUGAA) leads to the IPEX syndrome. Immunogenetics 53:435–439.
Castelo-Branco P, et al. 2004. Polypyrimidine tract binding protein modulates efficiency of polyadenylation. Mol Cell Biol. 24:4174–4183.
Castle JC. 2011. SNPs occur in regions with less genomic sequence conservation. PLoS One 6:e20660. doi: 10.1371/journal.pone.0020660
Chan SL, et al. 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. Genes Dev.
Charlesworth J, Eyre-Walker A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. Proc Natl Acad Sci U S A. 104:16992–16997.
Cheng Y, Miura RM, Tian B. 2006. Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics 22:2320–2325.
Colgan DF, Manley JL. 1997. Mechanism and regulation of mRNA polyadenylation. Genes Dev. 11:2755–2766.
Delahaye NF, et al. 2011. Alternatively spliced *NKp30* isoforms affect the prognosis of gastrointestinal stromal tumors. Nat Med. 17:700–707.
Denisov SV, et al. 2014. Weak negative and positive selection and the drift load at splice sites. Genome Biol Evol. 6:1437–1447.
di Giammartino DC, Nishida K, Manley JL. 2011. Mechanisms and consequences of alternative polyadenylation. Mol Cell. 43:853–866.
Force A, et al. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.
Genomes Project C, et al. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.
Genomes Project C, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.
Graham RR, et al. 2007. Three functional variants of IFN regulatory factor 5 (*IRF5*) define risk and protective haplotypes for human lupus. Proc Natl Acad Sci U S A. 104:6758–6763.

Gupta I, et al. 2014. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. Mol Syst Biol. 10:719.
Gyawali S, et al. 2010. Association of a polyadenylation polymorphism in the serotonin transporter and panic disorder. Biol Psychiatry. 67:331–338.
Ji Z, Lee JY, Pan Z, Jiang B, Tian B. 2009. Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A. 106:7028–7033.
Kaida D, et al. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature 468:664–668.
Kern AD, Kondrashov FA. 2004. Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. Nat Genet. 36:1207–1212.
Kimura M, Ohta T. 1971. Theoretical aspects of population genetics. Princeton (NJ): Princeton University Press.
Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A. 99:14878–14883.
Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. Genome Biol. 3: RESEARCH0008.
Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2006. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. J Theor Biol. 240:616–626.
Kurmangaliyev YZ, Sutormin RA, Naumenko SA, Bazykin GA, Gelfand MS. 2013. Functional implications of splicing polymorphisms in the human genome. Hum Mol Genet. 22:3449–3459.
Landrum MJ, et al. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42:D980–D985.
Lee JY, Ji Z, Tian B. 2008. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3′-end of genes. Nucleic Acids Res. 36:5581–5590.
Lee JY, Yeh I, Park JY, Tian B. 2007. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. Nucleic Acids Res. 35:D165–D168.
Mapendano CK, Lykke-Andersen S, Kjems J, Bertrand E, Jensen TH. 2010. Crosstalk between mRNA 3′ end processing and transcription initiation. Mol Cell. 40:410–422.
McVean GA, Charlesworth B. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 155:929–944.
Mustonen V, Lassig M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet. 25:111–119.
Naumenko SA, Kondrashov AS, Bazykin GA. 2012. Fitness conferred by replaced amino acids declines with time. Biol Lett. 8:825–828.
Nunes NM, Li W, Tian B, Furger A. 2010. A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. EMBO J. 29:1523–1536.
Ohta T. 1992. The nearly neutral theory of molecular evolution. Ann Rev Ecol Syst. 23:263–286.
Proudfoot N. 1991. Poly(A) signals. Cell 64:671–674.
Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. Science 320:1643–1647.
Schmidt S, et al. 2008. Hypermutable non-synonymous sites are under stronger negative selection. PLoS Genet. 4:e1000281.
Sheets MD, Ogg SC, Wickens MP. 1990. Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. Nucleic Acids Res. 18:5799–5805.
Shepard PJ, et al. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA 17:761–772.
Shi Y, et al. 2009. Molecular architecture of the human pre-mRNA 3′ processing complex. Mol Cell. 33:365–376.
Shin JH, et al. 2007. IA-2 autoantibodies in incident type I diabetes patients are associated with a polyadenylation signal polymorphism in GIMAP5. Genes Immun. 8:503–512.
Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Singh P, et al. 2009. Global changes in processing of mRNA 3′ untranslated regions characterize clinically distinct cancer subtypes. Cancer Res. 69:9422–9430.
Spies N, Burge CB, Bartel DP. 2013. 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. Genome Res. 23:2078–2090.
Stacey SN, et al. 2011. A germline variant in the *TP53* polyadenylation signal confers cancer susceptibility. Nat Genet. 43:1098–1103.
Stecher G, et al. 2014. MEGA-MD: molecular evolutionary genetics analysis software with mutational diagnosis of amino acid variation. Bioinformatics 30:1305–1307.
Takagaki Y, Manley JL. 1998. Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. Mol Cell. 2:761–771.
Thomas LF, Saetrom P. 2012. Single nucleotide polymorphisms can create alternative polyadenylation signals and affect gene expression through loss of microRNA-regulation. PLoS Comput Biol. 8:e1002621.
Tian B, Hu J, Zhang H, Lutz CS. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. 33:201–212.
Welter D, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42:D1001–D1006.
Wiestner A, et al. 2007. Point mutations and genomic deletions in *CCND1* create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. Blood 109:4599–4606.
Wilkening S, et al. 2013. An efficient method for genome-wide polyadenylation site mapping and RNA quantification. Nucleic Acids Res. 41:e65.
Yang Y, Mariati, Ho SC Yap MG. 2009. Mutated polyadenylation signals for controlling expression levels of multiple genes in mammalian cells. Biotechnol Bioeng. 102:1152–1160.
Yao C, et al. 2012. Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. Proc Natl Acad Sci U S A. 109:18773–18778.
You L, et al. 2015. APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. Nucleic Acids Res. 43:D59–D67.

Associate editor: Bill Martin